

**Q 1.** Recall the Gauss-Southwell rule for basic descent methods that we saw in class:  $\mathbf{d}_k = -\nabla_{i_k} f(\mathbf{x}_k) \mathbf{e}_{i_k}$ , where  $i_k = \operatorname{argmax}_{1 \leq i \leq n} |\nabla_i f(\mathbf{x}_k)|$  and  $\mathbf{e}_{i_k}$  is the vector that has 0 in all coordinates except for  $i_k$ , where it equals 1 (it is the  $i_k^{\text{th}}$  standard basis vector). Same as in the class, we assume that  $f$  is  $L$ -smooth. Prove that there exists  $\alpha > 0$  such that the Gauss-Southwell rule applied for an appropriate step size  $\alpha_k$  satisfies:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\alpha}{2} \|\nabla f(\mathbf{x}_k)\|_2^2.$$

How would you choose  $\alpha_k$ ? What can you say about the convergence of this method (discuss all three cases we have covered in class: nonconvex and bounded below, convex, strongly convex)? [10pts]

$$\mathbf{d}_k = -\nabla_{i_k} f(\mathbf{x}_k) \mathbf{e}_{i_k} \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k [\nabla f(\mathbf{x}_k)]_{i_k} \mathbf{e}_{i_k}$$

Since  $f$  is  $L$ -smooth:

$$\begin{aligned} f(\mathbf{y}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{y} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}_k\|^2 \\ f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \alpha_k \langle \nabla f(\mathbf{x}_k), \nabla f(\mathbf{x}_k) \mathbf{e}_{i_k} \rangle + \frac{\alpha_k^2 L}{2} [\nabla f(\mathbf{x}_k)]_{i_k}^2 \\ &= f(\mathbf{x}_k) - \alpha_k [\nabla f(\mathbf{x}_k)]_{i_k}^2 + \alpha_k^2 \cdot \frac{L}{2} [\nabla f(\mathbf{x}_k)]_{i_k}^2 \\ &= f(\mathbf{x}_k) + [\nabla f(\mathbf{x}_k)]_{i_k}^2 \left( \frac{L}{2} \alpha_k^2 - \alpha_k \right) \\ &\leq f(\mathbf{x}_k) + \frac{1}{n} \|\nabla f(\mathbf{x}_k)\|_2^2 \left( -\alpha_k + \frac{L}{2} \alpha_k^2 \right) \end{aligned}$$

Since we want  $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\alpha}{2} \|\nabla f(\mathbf{x}_k)\|_2^2$

$$\begin{aligned} \text{we have: } \frac{1}{n} \left( -\alpha_k + \frac{L}{2} \alpha_k^2 \right) &\leq -\frac{\alpha}{2} \\ \frac{L}{2} \alpha_k^2 + \left( \frac{n}{2} - 1 \right) \alpha_k &\leq 0 \\ \alpha_k &\in (0, \frac{n-1}{L}) \end{aligned}$$

$$\text{Then } f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\alpha}{2} \|\nabla f(\mathbf{x}_k)\|_2^2$$

According to the conclusion from the class.

$$\text{non convex: } \min_{0 \leq i \leq k} \|\nabla f(\mathbf{x}_i)\| \leq \sqrt{\frac{2(f(\mathbf{x}_0) - f^*)}{2(k+1)}}$$

$$\text{convex: } f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \frac{2\|\mathbf{x}^* - \mathbf{x}_0\|^2}{2(k+2)}$$

$$\text{Strongly convex: } f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq (-m_2)^{k+1} \frac{\|\mathbf{x}^* - \mathbf{x}_0\|^2}{2_2}$$

**Q 2.** Exercise 8 from Chapter 3 in Recht-Wright. Notation from the exercise:  $\mathbf{x}_* = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$ . You don't need to worry about getting the same constants as stated there, being off by constant factors (up to 4) is fine. [15pts]

8. **Weakly convex optimization.** Let  $f$  be a convex function with  $L$ -Lipschitz gradients. Assume that we know the true optimal solution lies in a ball of radius  $R$  about zero. In this exercise, we will show that minimizing a nearby strongly convex function will quickly produce a solution that is an approximate minimizer of  $f$ . Consider running the gradient method on the function

$$f_\epsilon(x) = f(x) + \frac{\epsilon}{2R^2} \|x\|^2$$

initialized at some  $x_0$  with  $\|x_0\| \leq R$ .

- (a) Let  $x_*^{(\epsilon)}$  denote an optimal solution of  $f_\epsilon$ . Is  $x_*^{(\epsilon)}$  unique?
- (b) Prove that  $f(z) - f(x_*) \leq f_\epsilon(z) - f_\epsilon(x_*^{(\epsilon)}) + \frac{\epsilon}{2}$ .
- (c) Prove that for an appropriately chosen stepsize, the gradient method applied to  $f_\epsilon$  will find a solution such that

$$f_\epsilon(z) - f_\epsilon(x_*^{(\epsilon)}) \leq \frac{\epsilon}{2}$$

in at most

$$\frac{R^2 L}{\epsilon} \log \left( \frac{8R^2 L}{\epsilon} \right)$$

iterations. Find a constant stepsize that yields such a convergence rate.

(a) Claim:  $f_\epsilon(x)$  is a strongly convex function.

$$\forall x, y \quad f_\epsilon(y) - f_\epsilon(x) = f(y) - f(x) + \frac{\epsilon}{2R^2} \|y\|^2 - \frac{\epsilon}{2R} \|x\|^2$$

Since  $f$  is convex, according to Taylor's theorem,

$$\begin{aligned} f(y) - f(x) &\geq \langle \nabla f(x), y-x \rangle + \frac{\epsilon}{R^2} \|x-y\|^2 + \frac{\epsilon}{2R^2} \|y-x\|^2 \\ &= \langle \nabla f(x), y-x \rangle + \frac{\epsilon}{2R^2} \|y-x\|^2 \end{aligned}$$

$$\text{We have: } f_\epsilon(y) - f_\epsilon(x) \geq \langle \nabla f_\epsilon(x), y-x \rangle + \frac{m}{2} \|y-x\|^2 \quad \forall x, y. \quad m = \frac{\epsilon}{R^2}$$

Thus,  $f_\epsilon(x)$  is strongly convex.

then  $x_\epsilon^{(y)}$  is an optimal solution of  $f_\epsilon$ ,  $x_\epsilon^{(y)}$  is unique.

$$(b) \quad \text{Since } f(z) - f(z) = -\frac{\epsilon}{2R^2} \|z\|^2 \leq 0 \quad \text{①}$$

$$f_\epsilon(x_\epsilon^{(y)}) - f(x_\epsilon^{(y)}) \leq f_\epsilon(x_\epsilon^{(y)}) - f(x_\epsilon^{(y)}) = \frac{\epsilon}{2R^2} \|x_\epsilon^{(y)}\|^2 \leq \frac{\epsilon}{2} \quad \text{②}$$

$$\text{then: } \text{①} + \text{②: } f(z) - f(z) + f_\epsilon(x_\epsilon^{(y)}) - f(x_\epsilon^{(y)}) \leq \frac{\epsilon}{2}$$

$$f(z) - f(x_\epsilon^{(y)}) \leq f_\epsilon(z) - f_\epsilon(x_\epsilon^{(y)}) + \frac{\epsilon}{2}$$

(c) According to Strong convexity of  $f(x)$   
 $X_{k+1} = X_k - \lambda \nabla f(X_k)$

According to the result in the class:

$$f(X_k) - f(X^*) \leq (1-m_2)^k \frac{\|X^* - X_0\|_2^2}{2}$$

$$\leq (1-m_2)^k \frac{2R^2}{2}$$

WMT:  $f(X_k) - f(X^*) \leq \frac{\epsilon}{2}$

$$k \geq \frac{1}{2m} \log \left( \frac{4R^2}{2\epsilon/2} \right)$$

Since  $m = \frac{\epsilon}{R^2}$   $\|X_0 - X^*\|_2^2 \leq 4R^2$

$$k \geq \frac{R^2}{2\epsilon} \log \left( \frac{8R^2}{2\epsilon} \right)$$

take  $\lambda = \frac{1}{L}$

we have  $k \geq \frac{R^2 L}{\epsilon} \log \left( \frac{8R^2 L}{\epsilon} \right)$

Choose  $\lambda = \frac{1}{L}$  as constant step size,

**Q 3** (Bregman Divergence). Bregman divergence of a continuously differentiable function  $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$  is a function of two points defined by

$$D_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Equivalently, you can view Bregman divergence as the error in the first-order approximation of a function:

$$\psi(\mathbf{x}) = \psi(\mathbf{y}) + \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + D_\psi(\mathbf{x}, \mathbf{y}).$$

- (i) What is the Bregman divergence of a simple quadratic function  $\psi(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2$ , where  $\mathbf{x}_0 \in \mathbb{R}^n$  is a given point? [5pts]

- (ii) Given  $\mathbf{x}_0 \in \mathbb{R}^n$  and some continuously differentiable  $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ , what is the Bregman divergence of function  $\phi(\mathbf{x}) = \psi(\mathbf{x}) + \langle \mathbf{x}_0, \mathbf{x} \rangle$ ? [5pts]

- (iii) Use Part (ii) and the definition of Bregmann divergence to prove the following 3-point identity:

$$(\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n) : D_\psi(\mathbf{x}, \mathbf{y}) = D_\psi(\mathbf{z}, \mathbf{y}) + \langle \nabla \psi(\mathbf{z}) - \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{z} \rangle + D_\psi(\mathbf{x}, \mathbf{z}). \quad [5\text{pts}]$$

- (iv) Suppose I give you the following function:  $m_k(\mathbf{x}) = \sum_{i=0}^k a_i \psi_i(\mathbf{x})$ , where  $\psi_i(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}_i\|_2^2 + \langle \mathbf{b}_i, \mathbf{x} - \mathbf{x}_i \rangle$ , where  $\{a_i\}_{i \geq 0}$  is a sequence of positive reals and  $\{\mathbf{b}_i\}_{i=0}^k, \{\mathbf{x}_i\}_{i=0}^k$  are fixed vectors from  $\mathbb{R}^n$ . Define  $\mathbf{v}_k = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n} m_k(\mathbf{x})$  and  $A_k = \sum_{i=0}^k a_i$ . Using what you have proved so far, prove the following inequality:

$$(\forall \mathbf{x} \in \mathbb{R}^n) : m_{k+1}(\mathbf{x}) \geq m_k(\mathbf{v}_k) + a_{k+1} \psi_{k+1}(\mathbf{x}) + \frac{A_k}{2} \|\mathbf{x} - \mathbf{v}_k\|_2^2. \quad [5\text{pts}]$$

$$\begin{aligned} (\text{i}) \quad D_\psi(\mathbf{x}, \mathbf{y}) &= \frac{1}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2 - \frac{1}{2} \|\mathbf{y} - \mathbf{x}_0\|_2^2 - \langle \mathbf{y} - \mathbf{x}_0, \mathbf{x} - \mathbf{y} \rangle \\ &= \frac{1}{2} \|\mathbf{x}\|_2^2 + \frac{1}{2} \|\mathbf{x}_0\|_2^2 - \langle \mathbf{x}, \mathbf{x}_0 \rangle - \frac{1}{2} \|\mathbf{y}\|_2^2 - \frac{1}{2} \|\mathbf{x}_0\|_2^2 + \langle \mathbf{y}, \mathbf{x}_0 \rangle - \langle \mathbf{y}, \mathbf{x} \rangle \\ &\quad + \langle \mathbf{x}_0, \mathbf{x} \rangle + \|\mathbf{y}\|_2^2 - \langle \mathbf{x}_0, \mathbf{y} \rangle \\ &\approx \frac{1}{2} \|\mathbf{x}\|_2^2 - \frac{1}{2} \|\mathbf{y}\|_2^2 - \langle \mathbf{y}, \mathbf{x} \rangle + \|\mathbf{y}\|_2^2 = \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \end{aligned}$$

$$(\text{ii}) \quad \phi(\mathbf{x}) = \psi(\mathbf{x}) + \langle \mathbf{x}_0, \mathbf{x} \rangle$$

$$\begin{aligned} D_\phi(\mathbf{x}, \mathbf{y}) &= \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla \phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = \psi(\mathbf{x}) - \psi(\mathbf{y}) + \langle \mathbf{x}_0, \mathbf{x} - \mathbf{y} \rangle - \langle \nabla \psi(\mathbf{y}), \mathbf{x}_0, \mathbf{x} - \mathbf{y} \rangle \\ &= \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = D_\psi(\mathbf{x}, \mathbf{y}) \end{aligned}$$

$$(\text{iii}) \quad D_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

$$\begin{aligned} D_\psi(\mathbf{x}, \mathbf{y}) &+ \langle \nabla \psi(\mathbf{z}) - \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{z} \rangle + D_\psi(\mathbf{x}, \mathbf{z}) \\ &= \psi(\mathbf{z}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{z} - \mathbf{y} \rangle + \psi(\mathbf{x}) - \psi(\mathbf{z}) - \langle \nabla \psi(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \\ &\quad + \langle \nabla \psi(\mathbf{z}) - \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{z} \rangle \\ &= \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \end{aligned}$$

$$\text{then we have } D_\psi(\mathbf{x}, \mathbf{y}) = D_\psi(\mathbf{x}, \mathbf{z}) + \langle \nabla \psi(\mathbf{z}) - \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{z} \rangle + D_\psi(\mathbf{x}, \mathbf{y})$$

$$(iv) \quad \text{Since } m_k(x) = m_{k+1}(x) - \alpha_{k+1} \psi_{k+1}(x)$$

WRT.  $m_k(x) \geq m_k(v_k) + \frac{\beta_k}{2} \|x - v_k\|_2^2$  ①

$$\psi_i(x) = \frac{1}{2} \|x - x_i\|_2^2 + \langle b_i, x \rangle - \langle b_i, x_i \rangle$$

$$\text{According to (ii): } D\psi_i(x, y) = \frac{1}{2} \|x - y\|_2^2$$

$$\begin{aligned} Dm_k(x, v_k) &= m_k(x) - m_k(v_k) - \langle \nabla m_k(v_k), x - v_k \rangle \\ &= \sum_{i=0}^k \alpha_i \left\{ \psi_i(x) - \psi_i(v_k) - \langle \nabla \psi_i(v_k), x - v_k \rangle \right\} \\ &= \sum_{i=0}^k \alpha_i D\psi_i(x, y) = \frac{\beta_k}{2} \|x - v_k\|_2^2 \end{aligned} \quad (2)$$

$$\text{Since } \psi_i(x) \text{ is convex, } m_k(x) = \sum_{i=0}^k \alpha_i \psi_i(x) \quad (\alpha_i > 0 \quad \forall i)$$

then  $m_k(x)$  is convex,

$$\text{Since } v_k = \underset{x \in R^n}{\operatorname{argmin}} m_k(x)$$

$$\text{then } \nabla m_k(v_k) = 0$$

$$\begin{aligned} \text{Consider } Dm_k(x, v_k) &= m_k(x) - m_k(v_k) - \langle \nabla m_k(v_k), x - v_k \rangle \\ &= m_k(x) - m_k(v_k) \end{aligned} \quad (3)$$

According to ② and ③, we have:

$$m_k(x) - m_k(v_k) = \frac{\beta_k}{2} \|x - v_k\|_2^2$$

$$\text{then: } m_{k+1}(x) \geq m_k(v_k) + \alpha_{k+1} \psi_{k+1}(x) + \frac{\beta_k}{2} \|x - v_k\|_2^2$$

**Q 4.** In class, we have analyzed the following variant of Nesterov's method for  $L$ -smooth convex optimization:

$$\begin{aligned}\mathbf{x}_k &= \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \mathbf{v}_{k-1} \\ \mathbf{v}_k &= \mathbf{v}_{k-1} - a_k \nabla f(\mathbf{x}_k) / L \\ \mathbf{y}_k &= \mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k),\end{aligned}$$

where  $L$  is the smoothness constant of  $f$ ,  $a_0 = A_0 = 1$ ,  $\frac{a_k^2}{A_k} = 1$ ,  $A_k = \sum_{i=0}^k a_i$ . We take  $\mathbf{x}_0 \in \mathbb{R}^n$  to be an arbitrary initial point and  $\mathbf{y}_0 = \mathbf{v}_0 = \mathbf{x}_0 - \nabla f(\mathbf{x}_0) / L$ .

Prove that we can state Nesterov's method in the following equivalent form:

$$\begin{aligned}\mathbf{x}_k &= \mathbf{y}_{k-1} + \frac{a_k}{A_k} \left( \frac{\frac{A_{k-1}}{a_{k-1}} - 1}{\frac{a_k}{A_k}} \right) (\mathbf{y}_{k-1} - \mathbf{y}_{k-2}), \\ \mathbf{y}_k &= \mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k).\end{aligned}\tag{1}$$

**Hint:** It is helpful to first prove that  $\mathbf{y}_k = \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \mathbf{v}_k$ .

[10pts]

$$\text{if, Claim ①: } \mathbf{y}_k = \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \mathbf{v}_k$$

$$\mathbf{y}_k = \mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k) = \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \mathbf{v}_{k-1} - \frac{1}{L} \nabla f(\mathbf{x}_k)$$

$$\text{Since } \mathbf{v}_k - \mathbf{v}_{k-1} = -a_k \cdot \frac{1}{L} \nabla f(\mathbf{x}_k)$$

$$\frac{a_k}{A_k} (\mathbf{v}_k - \mathbf{v}_{k-1}) = -\frac{a_k^2}{A_k} \cdot \frac{1}{L} \nabla f(\mathbf{x}_k) = -\frac{1}{L} \nabla f(\mathbf{x}_k)$$

$$\text{Thus: } \mathbf{y}_k = \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \mathbf{v}_{k-1} + \frac{a_k}{A_k} (\mathbf{v}_k - \mathbf{v}_{k-1}) = \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \mathbf{v}_k \quad \blacksquare$$

$$\begin{aligned}\text{Then, } \mathbf{x}_k &= \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \mathbf{v}_{k-1} \\ &= \mathbf{y}_{k-1} + \frac{a_k}{A_k} (\mathbf{v}_{k-1} - \mathbf{y}_{k-1})\end{aligned}$$

$$\text{Claim ②: } \mathbf{v}_{k-1} - \mathbf{y}_{k-1} = \left( \frac{A_{k-1}}{a_{k-1}} - 1 \right) (\mathbf{y}_{k-1} - \mathbf{y}_{k-2})$$

$$\text{According to claim ①: } \mathbf{y}_{k-1} - \frac{A_{k-2}}{A_{k-1}} \mathbf{y}_{k-2} = \frac{a_{k-1}}{A_{k-1}} \mathbf{v}_{k-1}$$

$$\Leftrightarrow \frac{A_{k-1}}{\alpha_{k-1}} y_{k-1} - \frac{A_{k-2}}{\alpha_{k-1}} y_{k-2} > v_{k-1}$$

$$v_{k-1} - y_{k-1} = \left( \frac{A_{k-1}}{\alpha_{k-1}} - 1 \right) y_{k-1} - \frac{A_{k-2}}{\alpha_{k-1}} y_{k-2}$$

$$= \left( \frac{A_{k-1}}{\alpha_{k-1}} - 1 \right) y_{k-1} - \left( \frac{A_{k-1}}{\alpha_{k-1}} - 1 \right) y_{k-2}$$

$$= \left( \frac{A_{k-1}}{\alpha_{k-1}} - 1 \right) (y_{k-1} - y_{k-2})$$

四

then:

$$\left\{ \begin{array}{l} x_k = y_{k-1} + \frac{\alpha_k}{A_k} (v_{k-1} - y_{k-1}) \\ = y_{k-1} + \frac{\alpha_k}{A_k} \left( \frac{A_{k-1}}{\alpha_{k-1}} - 1 \right) (y_{k-1} - y_{k-2}) \\ y_k = x_k - \frac{1}{L} \nabla f(x_k) \end{array} \right.$$